NOT TO BE FURTHER COPIED OR REPRODUCED

[* y VS x *]

15

Fitting, Insulation, Transparency

	Exh.	Page
Introduction 15-1	a	
15A. Lessons about fitting in general 15-1	1[b]	15-8
15B. Insulation and transparency 15-5	2[c]	15-13
insulated, at 15-5	3	15-15
transparent, at 15-6		
	h	
15C. Ways of thought and orders of fitting 15-7		
- straight lines can be simple 15-11		
- bends are more complicated 15-14		
* * *		
15H. How far have we come? 15-17		
15P. Problems		

15

Fitting, Insulation, Transparency

In the earlier chapters devoted to y vs. x we have dealt with the simplest situations. It is time to re-examine what we have done, and see what we can learn that applies to wider classes of fitting problems.

Fitting is the most powerful of all our tools for data analysis. Equally, it is still, as Cochran once said, the tool least satisfactorily taught.

One of the aims of this whole account is to prepare ourselves to study fitting seriously, and with enough background to appreciate the major issues. Every chapter helps to serve this purpose -- as well as to help us learn to analyze data of relatively simple structure. Some chapters are more obviously and more completely devoted to a background for fitting, as it will arise both in later chapters and beyond this account's scope. This chapter is one such.

15A. Lessons about fitting in general

In chapter 14 we went further into the details of fitting. As a result we came upon examples of some very general characteristics of all known fitting processes - characteristics that could only be avoided by the use of extrasensory perception or some similar miracle.

When we come to fit, whether our purpose is to assess or to clear out of the way, we yearn to choose our procedure

so as to listen only to the sort of message we are seeking. Arithmetic can never do this for us. The best that we are able to do is to make our procedure deaf to certain specific sorts - a few, not many - of the messages that we do not wish to hear.

Consider a man who is concerned with the slope of y as a function of x - a man who, as "slope" suggests, is thinking of y depending on x in a straight-line way. What can he do to listen only to straight-line messages? There is at least one thing he can do - one thing we all do, usually without conscious thought. He can avoid completely any effect of the general level - the general size - of y.

In section 10A we took

$$\frac{y_H - y_L}{x_H - x_L}$$

as our measure of slope. If each and every individual y were increased by, say, 365, both y_H and y_L would be increased by 365 and, since $y_H - y_L$ would not change, the fitted tilt would not change at all. To the message "all y's are increased (or decreased) by the same amount" our slope assessment is completely deaf. This is good; this can be done; not to do it would usually be close to shocking.

But what if we change the y's in a way that is not represented by a straight line? Suppose we change the y's in proportion to $(x - \min x)^2$ where min x is the smallest x represented in the data? This change is far from a straight-line

change. A purist looking for straight-line slopes might wish not to see it. But, since y's for high x's are certainly increased more than y's for low x's, the tilt we assess using our usual approach -- or any reasonable approach -- will surely become more positive after this change.

There is nothing unique about x^2, we could have said the same thing about any steadily (or jerkily) increasing function of x. An assessor of slope cannot avoid being sensitive to very many sorts of changes. Only a few special changes will leave its result unaffected. (This is one of the reasons why we speak of a <u>tilt</u> assessor rather than of a <u>slope</u> assessor.)

One of the arts of careful fitting is arranging that among the few changes that leave the result unaffected are the particular ones whose possible influence on the result we most wish to avoid.

The assessment of bend offers another example. There we took the following steps:

- first find, for each of two different intervals, the tilt of y against x and the tilt of $(x-c)^2 + d$ against x.
- then find the differences from interval to interval for each of these tilts and form the ratio of these differences in tilt.

The whole shape of this calculation is designed to reject what we most want to reject: any straight-line (= wholly unbent) behavior. A constant contribution to all y's disappears at the first step, where we take the tilts. A steady

rate of increase or decrease of y with x sneaks through the first step -- raising or lowering the two tilts of y against x by exactly equal amounts -- but is caught, and rejected, at the second step -- where the difference of these two tilts is taken.

Almost any other dependence of y on x will, however, provoke some response in the result given by this bend assessor. Care allows us to eliminate responses to a few sorts of change. Nature insists upon responses to almost all others.

* back to batches *

The same relation of the attainable to the desirable was already present when we worked with batches in chapter 4. Once we both admit that there may be "wild" values and decide that we should like to neglect them, it is easy to say what we would like to have, namely:

- a central value that is unaffected by a change in any ONE of the observations from which it is computed.

It is equally easy to see why it is impossible to have such a utopian central value. Consider increasing one y after another by an equal amount, say by 100 units, until all have been increased. Each step -- since only one value changes -- would not change our utopian central value at all. The overall change, however, in which each and every y increases by 100, is one that must affect anything worthy of the name of central value. There can be no such utopian central value.

If we cannot have all we would like, how much can we have? Median, trimean, and midmean all have a large part of what we want, namely:

- if we change just one number more and more, the time comes when further change (of this one value in this one direction) has no effect on our typical value.

So much we can have; all we might dream of we cannot.

15B. Insulation and transparency

From time to time we need special words -- words to help us remember important ideas -- words that make it easier for us to be reminded again and again of these ideas.

The last section is devoted to examples of one idea

insulation

We have stressed that fitting procedures may be such as to give results that are unaffected by certain kinds of changes in the data. It is natural to say that either the fits or the fitting procedures are

insulated

from these changes.

Using the new term, we can rephrase the main message of the last section as follows:

- it is good to have a fitting procedure and its results insulated from changes we do not want to be bothered by.

- the changes that any one fitting procedure is insulated against are few and far between -- among all possible changes.

- choosing a fitting procedure that is insulated from the right changes is an important part of the fitter's art.

Another idea can help us, especially as our problems and techniques become a little more complicated. This is the idea of

transparency

the converse of the idea of insulation. If a fit is insulated from a change, NONE of the change shows up in that fit, it all goes elsewhere.

If a fit is

transparent

to a change, ALL the change goes into the fit. No effects can be found elsewhere -- in particular, the residuals are not changed at all.

We have seen many examples of transparency:

- when working with batches of numbers, adding the same constant to every number changes any of our central values by that same constant, so that the residuals are unchanged.

- when working with tilts, adding A+Bx to every y adds A+Bx to the fitted line, each residual is unchanged.

- when working with bends, adding $A+Bx+Cx^2$ to every y adds $A+Bx+Cx^2$ to the fitted quadratic, each residual is unchanged.

The idea of transparency is important because a transparent fit can be judged correct -- or in error -- by examining its residuals. We do not need to know what has been fitted. This is both a practical convenience and a meaningful principle.

The overall situation is shown schematically in exhibit 1.

For very different reasons, then, a fitting process is not disturbed by either:

- those changes against which it is insulated, because they do not affect its fit.
- those changes to which it is transparent, because they do not affect its residuals.

Only few kinds of changes fall into these two nice categories.

The vast majority of all kinds of changes are challenges to and difficulties for a fitting process because their effects appear BOTH in the fit AND in the residuals.

15C. <u>Ways of thought and orders of fitting</u>

If we change all our data according to

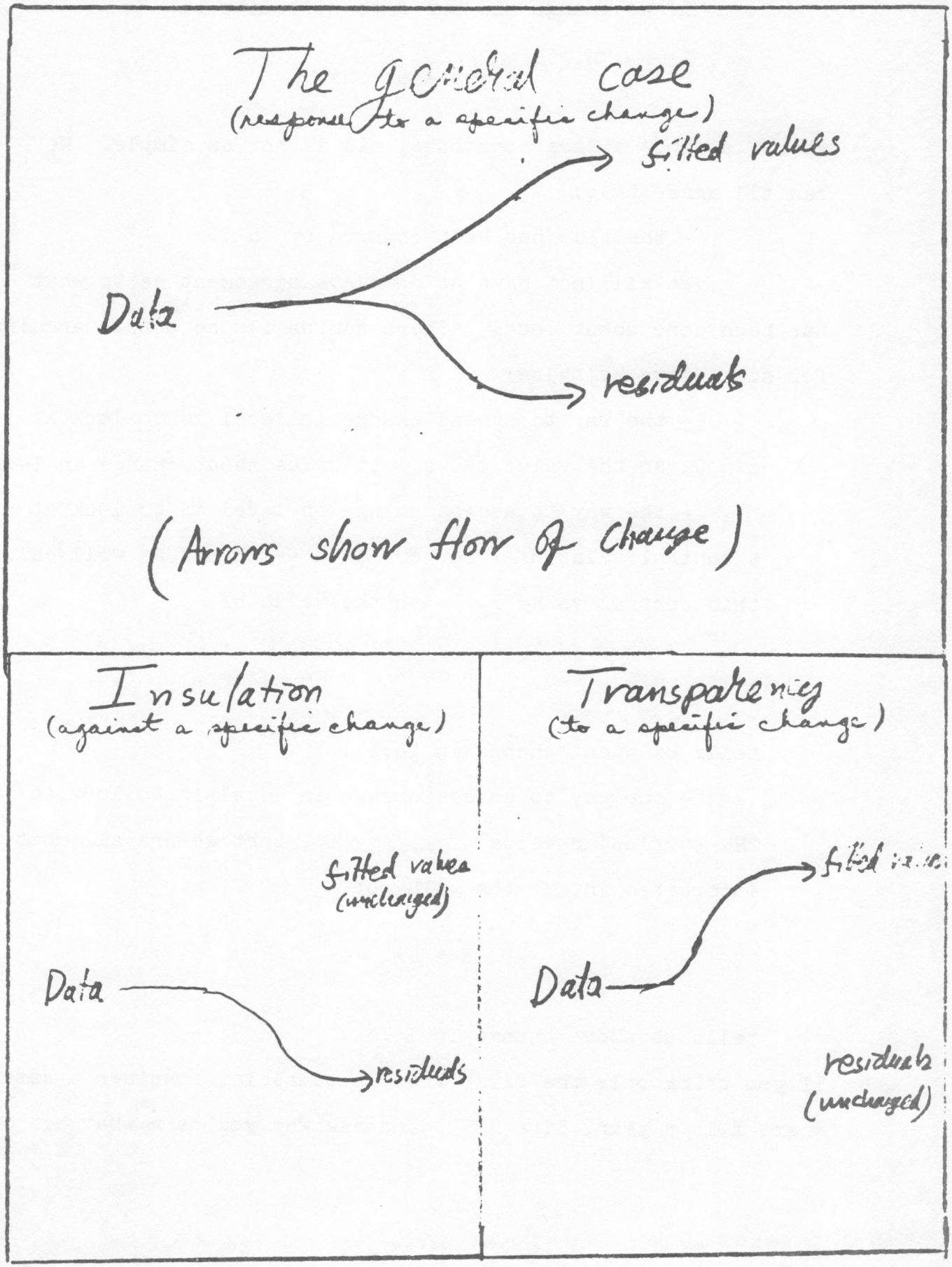
$$y \to y+c$$

where c is a constant, no one will challenge such statements as:

- we did not change anything but level.
- we changed level by c.

It is easy to talk about changes in level alone.

Schematic Views of the Effect of A Change in the Data on a Fitting Process



If we change all our data according to

$$y \to y + a + bx$$

where a and b are constants, all is not as simple. We can all agree that:

- the tilt has been changed by b .

We will not have as complete agreement as to what has been done about level. There can easily be people arguing for such views as these:

- the way to assess change in level is to look at $x = 0$, so the value of a tells us about change in level.

- the way to assess change in level is to look at a central value of x for our data -- we may as well call this central value x_C -- so the value of

$$a + bx_C$$

tells us about change in level.

- the way to assess change in level is to look at THE particular value of x, say x_Q, that we are all most interested in, so the value of

$$a + bx_Q$$

tells us about change in level.

If you think only the first view is sensible, consider a case where x is a year, like 1957, and ask why you care what a

straight line fitted to current data does somewhere between 1 B.C. and 1 A.D. If you think only the second view is sensible, consider a case where the only x's at which you can measure are far away on each side of the range of x's where the answer matters, and there are 10 times -- or 100 times -- as much data on one side than on the other. If you think that only the third view is sensible, consider what to do when three strong-minded experts disagree completely about the proper value of x_Q.

The easy way out is this:

- don't try to decide what change in level is involved in any change that includes a change in tilt.
- ask only what change in overall fit is required.

If we take this view, then the general pattern of our fitting procedure naturally takes this shape:

- first, find an $f+gx$ which makes

$$y - (f+gx)$$

untilted. We know about tilts and can forget about levels.

- second, find an h which puts

$$[y - (f+gx)] - h$$

at zero level. We aren't changing tilt, so we can think only about level, which we know how to do.

- finally, take

$$a+bx = [f+gx] + h$$

as the fitted line.

This sort of approach sneaks us neatly by the difficulty with the general questions:

What is level?

What is change in level?

by avoiding the first question and confining the second question to cases where level alone changes.

If this is to be the approach, what are the details? The first step is to fix a way to assess tilt that is:

- insulated from changes in level (we know what they are, when alone).
- transparent to changes in tilt (we know what they are, alone or combined with changes in level).

After we have used this assessor to let us subtract all the tilt we should, we can finish the fit by using any process we like that is

transparent to changes in level

In this latter step we can fit a pure change in level and avoid all worry about tilt.

* straight lines can be simple *

When we fit a straight line using three portions of the data each summarized by a central point defined by means, we follow this pattern exactly. First we assess a tilt from

$$b = \frac{y_H - y_L}{x_H - x_L}$$

and subtract

$$a+bx = b(x-d)$$

from each y, where a or d is chosen purely for numerical convenience. Then we assess a level from

$$c = \frac{1}{3}[(\text{adjusted } y)_H + (\text{adjusted } y)_M + (\text{adjusted } y)_L]$$

and subtract c from the result of the previous adjustment.

Here b is insulated from changes in level alone and transparent to changes in tilt. As for c , it is transparent to changes in level -- and we do not have to ask how changes in tilt affect it, since any change in tilt will have already been absorbed by a corresponding change in b before we come to calculating c.

Schematically, the process is as shown in exhibit 2. We manage to do a thorough job of keeping tilt and level out of the residuals, but only by never trying to say exactly what level is in the presence of tilt.

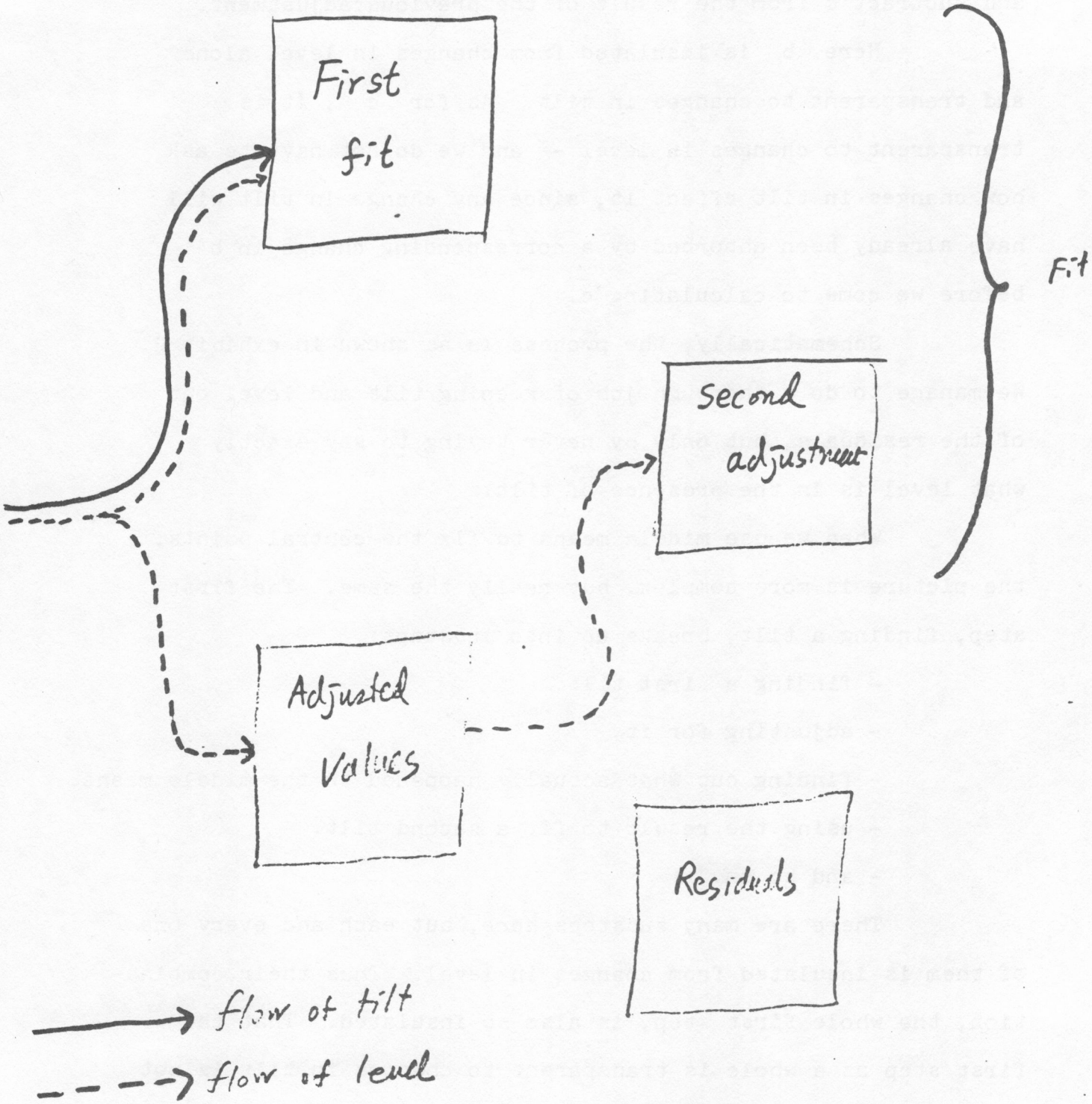
When we use middle means to fix the central points, the picture is more complex, but really the same. The first step, finding a tilt, breaks up into substeps:

- finding a first tilt.
- adjusting for it.
- finding out what actually happened to the middle means.
- using the result to fit a second tilt.
- and so on.

There are many substeps here, but each and every one of them is insulated from changes in level. Thus their combination, the whole first step, is also so insulated. That the first step as a whole is transparent to changes in tilt is not quite as obvious. In sensible practice, it is transparent,

exhibit 2 of chapter 15

Schematic Diagram of the Two-step Fitting of Tilt and Level



but there can be pathological -- sick, that is -- situations, where it need not be.

Let us look at such a sick situation. Exhibit 3 shows two sets of three points -- one set x's, the other o's -- which overlap from left to right. If we take these sets as two portions -- something any sensible fitter would shun like the plague -- and use medians to fix central points, we find that each of the lines in the exhibit makes

$$\text{median(adjusted y's)} = 0$$

for each portion separately. If we were so foolish as to use these two overlapping portions, there are two possible final fits, two different apparently satisfactory sets of residuals. If we change the data by

$$y \to y + a + bx$$

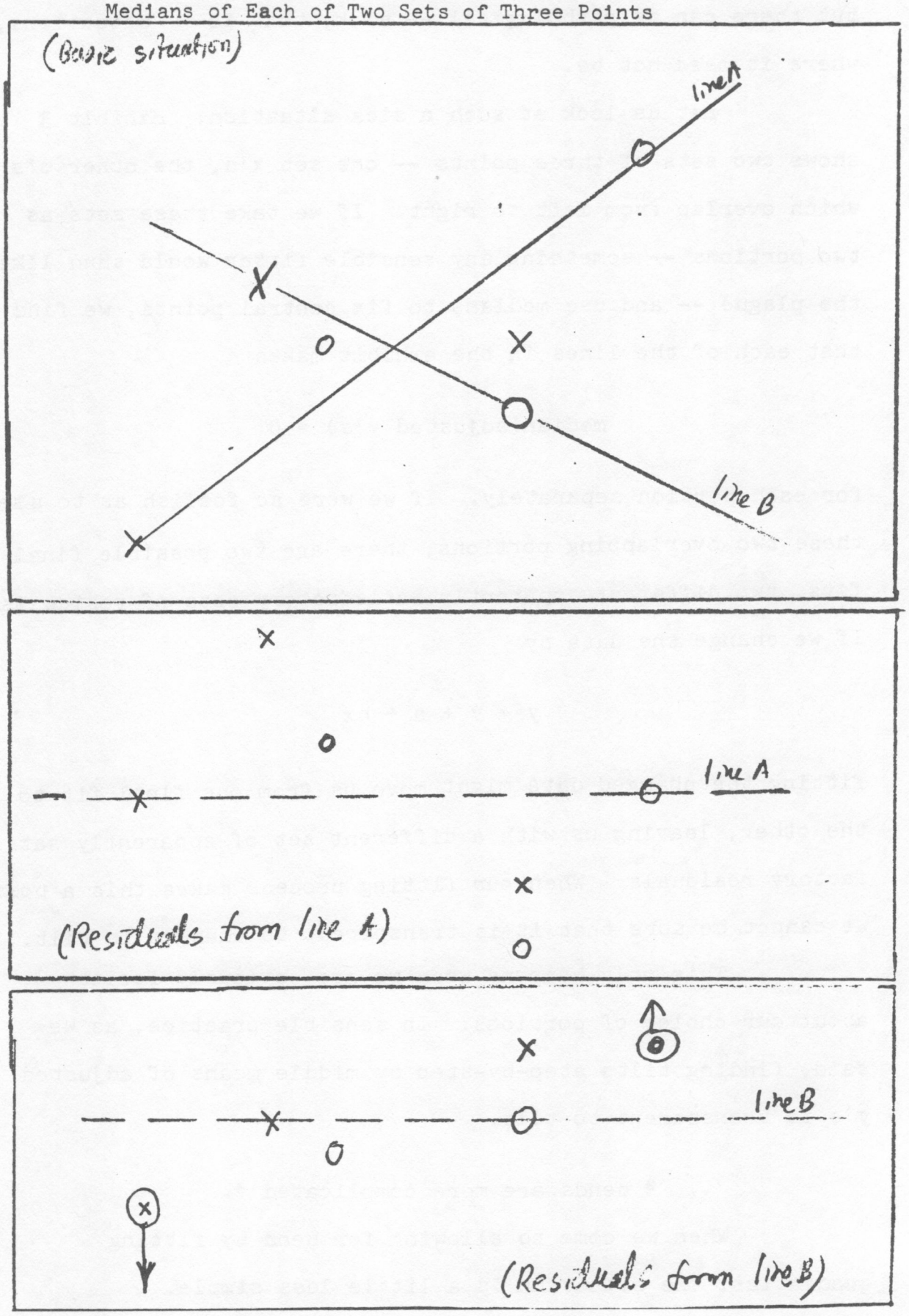
fitting the changed data might move us from one final fit to the other, leaving us with a different set of apparently satisfactory residuals. When our fitting process makes this a possibility, we cannot be sure that it is transparent to changes in tilt.

This only happens when we are obviously foolish about our choice of portions. In sensible practice, as we said, finding tilts step-by-step by middle means of adjusted y's is transparent to tilt.

* bends are more complicated *

When we come to allowing for bend by fitting quadratics, the situation is a little less simple.

A "sick" Example With Two Distinct Lines Through the Medians of Each of Two Sets of Three Points



If we were content to use means to summarize portions, all would be exactly as before. We use difference of tilts as a basis for choosing the coefficient of x^2. As we have said in other words, this difference is insulated from changes in tilt and level combined. We compare the difference in tilt for the data with the difference in tilt for x^2, and use the ratio to fix the coefficient of x^2. This makes the whole step transparent to changes in bend.

Now we are back to fitting tilt and level, and can proceed as before.

If, as we usually should, we use middle means to summarize the portions, things are almost -- but not quite -- the same. A change in tilt, if it is large enough, can change which data points provide the middle mean for a portion. As a result, the bend we assess is only

approximately insulated from changes in tilt

As a result, we do at least one of:

- subtracting a rough tilt before fitting a bend (as in section 14D).
- going back after fitting a bend and a tilt, and polishing both a little (as in section 14C).

Neither of these is difficult or really burdensome. Either is a little more work. Again all is well in sensible practice. We do have transparency for practical purposes.

*****

15H. How far have we come?

Having done enough fitting -- in earlier chapters -- to develop a good background, we have now gone on to consider how few changes in the data leave any particular aspect of a fit unaffected -- how few aspects of change the results of any particular fitting process are insulated from. As a result, choosing which changes are to have no effect is an important part of selecting a fitting process.

We also want to have most fitting processes transparent to those changes in the data that can be exactly fitted.

It is easy to be sure of how much a level changes when only level is changed. It is easy to be sure how much tilt changes when level and tilt both change. It is not easy -- on the other hand -- to be sure how much level changes when both tilt and level change together. This sort of fact influences the general structure of our fitting processes, leading us, for example, to fit tilt first, with little attention to level, and then fit a level to the adjusted values.

When fitting quadratics, this leads us to fit bend first, then tilt, and then level. To gain the advantages of middle means, we blur the simplicity of this a little by going back and polishing the fit of both bend and tilt once both have been well fitted. This is just a blurring -- needed because we have only approximate insulation -- not a major change in attack.

In general, we avoid asking about how much of a complicated change is of a simple sort -- how much a change by a+bx

changes level, or how much a change by $d+ex+fx^2$ changes either level or tilt. We can do a good job of finding fits and residuals without asking such questions.

15P. Problems

1. Select a batch of (x,y) points from the problems for chapter 8, 9, 10, or 11 and fit straight lines -- by arithmetic as in chapter 11 -- to (i) y against x, (ii) y + 432 against x, (iii) y + 1.2x against x, (iv) y + 1.2x + 432 against x. Compare the fitted formulas. State why they are related in the way that they are.

2. Trace the relationships of two of these fits through all the detailed stages of calculation.

3. An arithmetic fit of y - 1.23x + 456 against x leads to the fitted formula

$$\sim 12.34x - 57 .$$

What formula would be found if y were fitted against x?

4. If the fitting in problem 1 were done graphically, the relationships among the fits would be: -- exactly the same as -- almost the same as -- roughly the same as -- quite different from -- the relationships found in problem 1. Choose one of these answers and justify your choice.

5. Fitting y against x arithmetically gives

$$y \sim 432 - 98x + 3x^2 .$$

What will fitting $y - 3x + 2x^2$ against x give? What will y against x - 2 give?

6. Which of the results in the complete and detailed solutions of problems 1 to 5 (such solutions always take us as far as plotted residuals) illustrate transparency?

7. What results in the same complete and detailed solutions illustrate insulation?